# COMPUTATIONAL STUDIES TO INVESTIGATE HUMAN DNA REPAIR ASSOCIATED MALIGNANCIES WITH SPECIAL RELAVANCE TO COLORECTAL CANCER

**SYNOPSIS**

*Submitted in partial fulfillment of the requirements for the degree of*

**DOCTOR OF PHILOSOPHY**

IN

**BIOINFORMATICS**

By

**ANKITA**

Enrolment No. 136502

Under the supervision of

**Dr. Tiratha Raj Singh**

**DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS**

**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY**

**WAKNAGHAT, SOLAN-173234, HP, INDIA**

**April, 2018**

# COMPUTATIONAL STUDIES TO INVESTIGATE HUMAN DNA REPAIR ASSOCIATED MALIGNANCIES WITH SPECIAL RELAVANCE TO COLORECTAL CANCER

**ANKITA**

## INTRODUCTION

DNA repair is a collection of processes that help to identify and corrects damage that occur to the DNA molecules. The DNA repair ability of a cell is vital to the integrity of its genome and also for normal functioning of an organism [1]. DNA on daily basis faces thousands of damages through endogenous and exogenous damaging factors. These damaging factors include the formation of oxidative species via metabolism, base loss, radiations, and chemicals exposure that severely compromise its integrity [2]. These factors cause the genomic instability due to DNA damage, aberrant DNA replication or uncoordinated cell division that in most cases leads to chromosomal aberrations and gene mutations [3]. The maintenance of genome integrity is essential for organism survival and for the inheritance of traits to offspring. Though organisms need genome stability, they must allow some mutational changes to drive adaptation and evolution [4].

The genome contains all of the information required to build and maintain the organism that helps it to grow and develop. Therefore to deal with such damages there is a requirement of the mechanism that takes care of all forms of the alterations. By default, the cells have an inbuilt mechanism called DNA repair that take care of all occurring damages that happen to the DNA on daily basis [5]. Based on the type of DNA damage, the cell has evolved to gain a suitable method to repair the damages. These repair methods have been divided into two general classes that involves direct reversal method and damaged bases removal [6]. The damaged bases are further compensated by a different set of mechanisms depending on the degree and types of DNA damage which includes, base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), and double-strand break repair (DSBR) [6]. The factors affecting the DNA and their consequences along with the repair type are illustrated in Figure 1.

Knowledge of the human genome helped to determine the genetic cause of diseases, as some of the rare diseases are caused by the change in a variant. Today the world is fighting with

the diverse form of complex diseases that itself poses a challenge to overcome them. The varieties of cancers are there in the society that needs a quick measure to reduce the load of the patient's sufferings. In cancerous condition, major genomic alterations have been found in the tumor cells in comparison to the healthy cells [7]. Such alterations have been detected through the comparison of the tumor cells with the normal ones. The cancer cells often lack a method of DNA repair, and this deficit in most cases leads to tumorigenesis.

There are a variety of cancers that are present to date such as lung cancer, breast cancer, colorectal cancer, prostate cancer, skin cancer, endometrial cancer, bladder cancer, kidney cancer, thyroid cancer, uterine cancer, pancreatic cancer, oral cancer etc.[8]. The lung cancer, breast cancer, and colorectal cancer (CRC) are the top three cancers majorly responsible for causing the death of people worldwide. The constant increase in CRC cases in India is a matter of concern and requires effective therapies to control it [9].

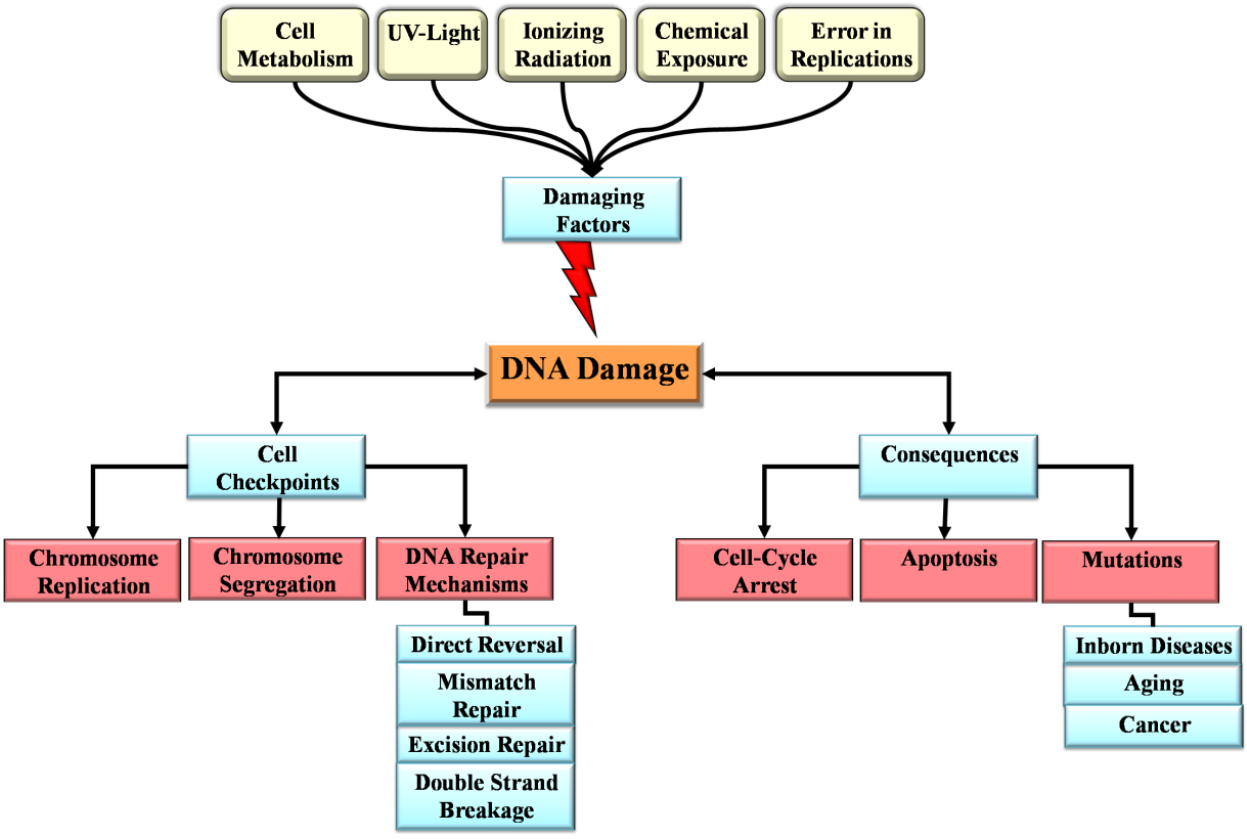


**Figure 1.** DNA Damage responses and their integrity in genome homeostasis.

Our study targeted the Lynch syndrome and associated cancer predominantly CRC. Lynch syndrome, often called hereditary nonpolyposis colorectal cancer (HNPCC), is an autosomal

inherited disorder that boosts the risk of multiple types of cancer, but, primarily cancer of the colon and rectum (i.e. CRC) [10]. The syndrome shows an autosomal dominant pattern and it depicts that even a single inherited copy of the altered gene is sufficient to amplify the cancer risk. This syndrome initiates due to faulty mismatch repair mechanism [11]. Other cancer risk involves that of the endometrium, stomach, ovary, small bowel, hepatobiliary tract, urinary tract, brain, and skin [12]. The common genetic variations such as MLH1, MSH2, MSH6, PMS2, and EPCAM increase the risk of developing Lynch syndrome [13]. CRC that is also known as cancer of bowel, colon, or rectum affects the colon and the rectum. In united states about 1 in 21 men and 1 in 23 women have the tendency to develop CRC during their lifetime [14]. In India, the age-standardized rates of CRC have been estimated to be 4.2/100,000 for males and 3.2/100,000 for females respectively. India is amongst the most heavily populated country but as their economies have developed the incidences for CRC has shown to be increased [15]. It has been noticed that majority of cases goes unnoticed and therefore shows poor statistics in terms of CRC cases.

The carcinogenesis mechanism follows three major mechanisms i.e. chromosomal instability (CIN) that accounts for ~85% CRC cases, microsatellite instability (MSI) accounts for ~15% CRC cases, and CpG island methylator phenotype (CIMP) ~5% CRC cases [16]. Along with the DNA repair pathways; there are multiple pathways that play role in cancer progression which involves Wnt, Hedgehog, TGFβ, MAPK, mTOR, JAK/STATs, COX, VEGF, NF-kB, and Notch signaling. The key genes of the pathway, the type of repair mechanism it affects, damaging factors, and the role played by pathways have mentioned in Table 1. The study implicated the Wnt and TGFβ study specifically as these pathways have been found to be the major targets of CRC.

To analyze the data in the proposed study, various bioinformatics and systems biology approaches along with various statistical measures were utilized to determine the structural and functional impacts of various biological entities. Some new gene regulators have been identified for the genetic disorders and the cancer conditions that provide clues about putative predicted biomarkers for these diseases. The bottom-up and top-down approaches were implemented for annotating the complex biological networks. The annotation and functional enrichment were made through various pathways and network motifs to generate the biological knowledge. Our

overall study is focused on the role of DNA repair mechanisms in the Lynch syndrome-associated cancers. Based on the worldwide limitations in research area for CRC and the Lynch syndrome associated cancer (specifically CRC, and cancer of endometrium), the study has been performed with the following objectives:

**OBJECTIVES:**

- **Development of a Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies: DREMECELS**
- **Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer (CRC).**
- **Study of Network and Structure Based Inference of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Impact on Colorectal Cancer.**

**Table 1.** Damaging effect on cellular signaling and its influence on DNA repair signaling pathways

| Signaling Pathways | Effected Repair Mechanism | Key Genes | Factors | Role |
|---|---|---|---|---|
| Wnt signaling | Double strand break repair (DSBR) NHEJ, HRR | Wnt/β-catenin | oxidative stress | gene regulation, cell polarity, adhesion, maintaining cell-behavior |
| Hedgehog signaling | BER, MMR, NER, NHEJ, HRR | Hh/GLI | endogenous molecules, oxysterols | embryonic development, carcinogenesis, maintaining cancer stem cells (CSCs), mesenchymal transition (EMT) |
| TGFβ signaling | NHEJ, HRR | TGFβ, receptors I and II, SMADs | endogenous DNA damage, radiations | cell proliferation, tissue homeostasis |
| MAPK signaling | NHEJ, HRR | MK2, MAPKKK, MEKK | replicative stress, reactive oxygen species | progression of the replication fork |
| mTOR signaling | Direct reversal, NER | CHK1 | endogenous DNA damage, replication stress | cell metabolism, growth, proliferation, survival |
| JAK/STATs signaling | NHEJ, HRR | JAK, STAT | cytostatic drugs, radiations | heterochromatin formation, development, homeostasis |
| COX signaling | NER | COX-2 | UV-light | haematopoietic stem cell homeostasis, suppression of embryonic stem cell apoptosis |
| VEGF signaling | Direct reversal, excision repair mechanisms | VEGF-A | endogenous DNA damage | cell homeostasis |
| NF-kB signaling | NHEJ, HRR | NF-kB | metabolic and exogenous sources | DNA transcription, cytokine production, cell survival |
| Notch signaling | NHEJ, HRR | ATM | reactive oxygen species | cell-fate, homeostasis |

## OBJECTIVE - I

**Development of a Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies: DREMECELS**

## METHODOLOGY

DREMECELS comprises manually collected data through a variety of standard resources and by performing an extensive literature search to retrieve genes implicated in cancers and DNA repair mechanisms. Events like gene interactions, enrichment, drug sensitivity, and methylation have added substantial credence to the database. The overall architecture of the database is given below; it represents four steps; the input data, the analysis tools (MSA, BLAST) [17, 18], processing phase, output generation phase, and finally the development of a GUI-based application (Figure 2).

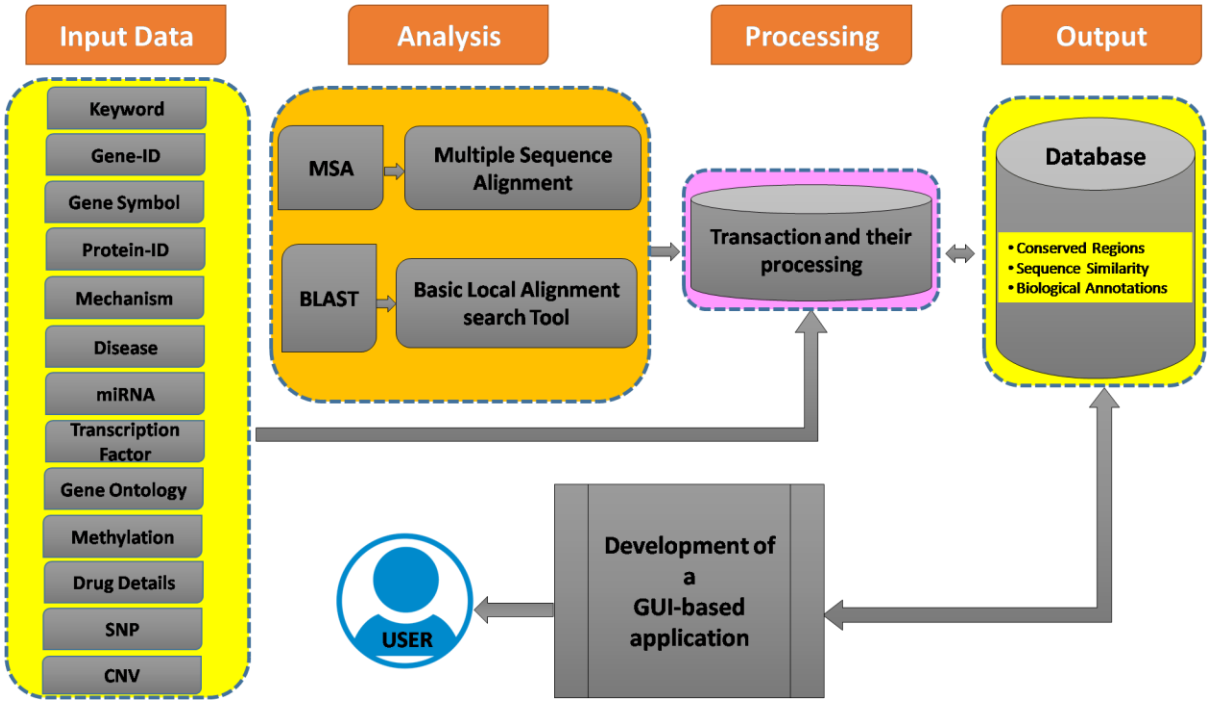


**Figure 2.** The comprehensive architecture of the DREMECELS.

### Database Configuration

The database follows relational database system as it integrates the data through MYSQL implementation. Programming languages like HTML, CSS, JavaScript, and PHP were used for

developing a graphical user interface (GUI). The connection between the interface and the back-end was established using phpMyAdmin being hosted on apache server.

## RESULTS AND DISCUSSION

DREMECELS comprises of 156 DNA repair (MMR and BER related) gene biomarkers associated with the CRC, EC, and LS. The database offers an easy to use and efficient method to search and retrieve data of associated genes along with other significant genomic and proteomic parameters. The overall percentage of genes falling under the category of MMR is 48%, approximately 23% were found to be involved in BER, and the genes falling under both categories contribute 29% of total genes in the database (Figure 3).

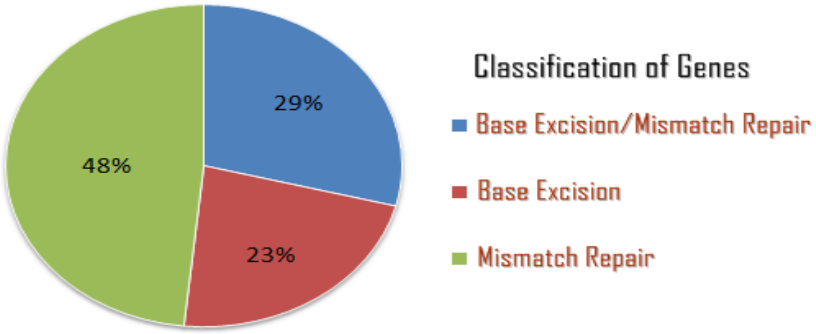


**Figure 3.** DNA repair genes and associated mechanisms; the percentage (%) shows the number of genes present in each mechanism.

From query search option, the search can be performed and will redirect to the results page if the search query matches otherwise return the error message *"No such record found"*. The results page will display following outputs:

- The comprehensive information in a tabular format covering gene ID, gene symbol, protein ID, protein name, mechanism, annotation, transcription factor, pathways, disease, PubMed ID, miRNA, interaction ID, PDB ID, ENSEMBL ID, OMIM ID, conserved domains that are directly linked to the gene card [19-22].
- The pictographic illustration of the interacting genes through embedded STRING page and overall connectivity to the original resource [23].
- The sequence conservation was computed through MUSCLE, thus represented in the form of the aligned sequences signifying the level of similarity.

- The methylation data include gene ID, chromosome number, start and end positions, gene methylation (%), technology, disease type, experiment ID, and CpG number [24].
- The drug details include manually created drug ID, drug name, tissue name, disease, cell line, IC-50, target gene, PubChem ID, and data source [25].
- The copy number variation (CNV) details include the sample, gene, expression, CN_Type, copy number, copy number segment position, and disease [26].
- The SNP detail includes gene, transcript, sample name, amino acid mutation, somatic mutations for substitution-missense, substitution-intronic, substitution-nonsense, substitution-coding silent, deletion-frameshift and deletion-in frame and disease [26, 27].

Also, we have integrated BLAST [18] in our database for performing similarity search for identifying close homologs. The method employs user-defined protein in a fasta format as a query search against the sequences available in the protein archive of our database. The main usage of deploying BLAST tool in this portal is to characterize hypothetical sequences related to the BER, and MMR mechanisms and the disease implicated. It results in the homologous sequences from the database on the basis of sequence similarity scores that is counted through E-value. We have implicated the standard parameters for protein BLAST to make the search smooth and effective.

## CONCLUSION

DREMECELS is a first-ever resource predominantly for colorectal, endometrial cancers and both cancers allied to Lynch syndrome. It serves as an extensive compilation for BER, and MMR genes comprising of numerous genetic parameters such as comprehensive gene-details, methylation, pathways, diseases, mutations etc. All data has been collected manually and is supported by literature references for these malignancies that provide a knowledge-based resource and thus allow researchers and clinicians to have a biological overview of the genes implicated in these diseases. The database would help save time and efforts of researchers involved in the field through easy accessibility to data, and thus will facilitate in biological discoveries. The database will be updated on a regular basis to keep state-of-the-art information to the academicians and researchers. The database is organized to provide clarity, ease of access, download and fast browsing capability.

## OBJECTIVE - II

## Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer (CRC)

### METHODOLOGY

The Wnt model was developed by considering the Sivakumar et al. as a reference model to study the biological signaling processes. The pathway model is imported and analysis is performed through MATLAB toolbox, SimBiology [28]. It provides programmatic tools to model, simulate, and analyze dynamic systems focusing principally on biological systems. The overall step-wise description of the method followed is given in Figure 4. The method comprises of two forms of analysis; first one is based on quantitative simulations on pathways and the other is for network motif determination to find the putative candidates as regulatory elements.

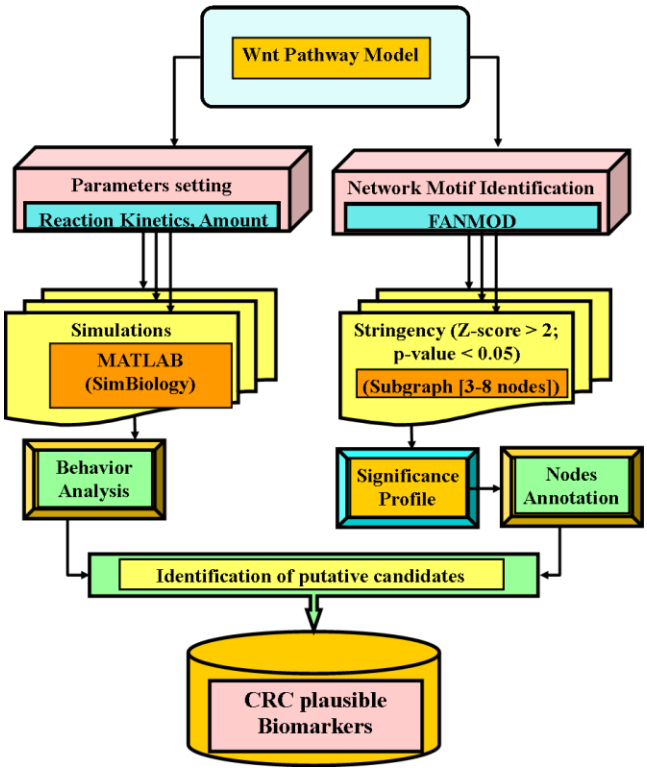


**Figure 4.** The flowchart of the methodology followed for pathway analysis.

### RESULTS AND DISCUSSION

#### Pathway Analysis

The pathway simulations were performed for all the entities by changing the parameters for every set of experiment. The solver Ode45 (Dormand-Prince) is used to run the simulations; the solver provides an explicit method to solve set of ODEs with minimum error rate [29]. Different parameters were implicated depending on various standards provided in the literature (Table 2). The concentrations of the model have obtained via the SBMLsqueezer; a CellDesigner plugin that generates reaction kinetics for the biochemical network [30]. In each reaction mechanism, the kinetic equation is derived from its stoichiometry, regulatory mechanisms and the involved species (such as simple molecules, proteins etc.). The rate laws are derived by considering individual reaction for every set of participating substrates, products, and the regulators.

**Table 2.** Quantitative parameters for all the analyzed species in simulation studies with varied concentrations

| Type | Name | Location | Concentrations | Solver |
|---|---|---|---|---|
| Species | β-Catenin | Plasma Membrane | 5 | Ode45 (Dormand-Prince) |
| Complex | APC, β-Catenin, GSK3β, Axin, PP2A, Diversin, CK1 | Plasma Membrane | 5 | |
| Complex | APC, Axin, PP2A, Diversin, CK1, β-Catenin, β- TrCP, GSK3β | Plasma Membrane | 0 | |
| Complex | APC, Axin, PP2A, Diversin, CK1, β-Catenin, β- TrCP, β- TrCP, GSK3β | Plasma Membrane | 5 | |
| Complex | Complex_br_(Wnt/Frizzled) | Plasma Membrane | 0 | Ode45 (Dormand-Prince) |
| Species | β-Catenin | Plasma-Membrane | 5 | |
| Species | β-Catenin | Nucleus | 0 | |
| Complex | TCF, Smad4, β-Catenin | Nucleus | 0 | |
| Species | β-Catenin | Nucleus | 0.53 | |
| Complex | APC, Axin, PP2A, Diversin, CK1, β-Catenin, β- TrCP, GSK3β | Nucleus | 0 | |
| Species | β-Catenin | Nucleus | 5 | |
| Complex | APC, Axin, , PP2A, Diversin, CK1, β-Catenin, β- TrCP, GSK3β | Nucleus | 5 | |
| Species | β-Catenin | Nucleus | 6 | |

Initial simulation was run for all set of components of the pathway; however no specific behavior was analyzed. Next run was performed for β-catenin (plasma membrane) with concentration of 5 micromolar (μM), the complex (APC, β-catenin, GSK3β, Axin, PP2A, Diversin, CK1) with 5 μM concentration and the complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β) considering to be at 0 μM. The results have shown that even though there is high amount of β-catenin but as long as it's captured by the destruction complex, β-catenin slows down its activity and its concentrations diminishes from its initial amount of 5 μM to 0.2 μM. This illustrates the fact even with the high concentration β-catenin will not work until the

destruction complex is there and it will regulate the level of β-catenin irrespective of its own higher or lower concentrations as shown in the graph at a time it is having a concentration of 5 μM and 0 μM (Figure 5). Complex-1 and β-catenin have same behavior while complex-2 is showing the opposite behavior (Figure 5a). However, all three reaching at steady state at the almost same time and this reflects that initial few seconds (i.e. 3-4 milliseconds) are crucial for their activity to finally reach and follow the steady state (Figure 5b).

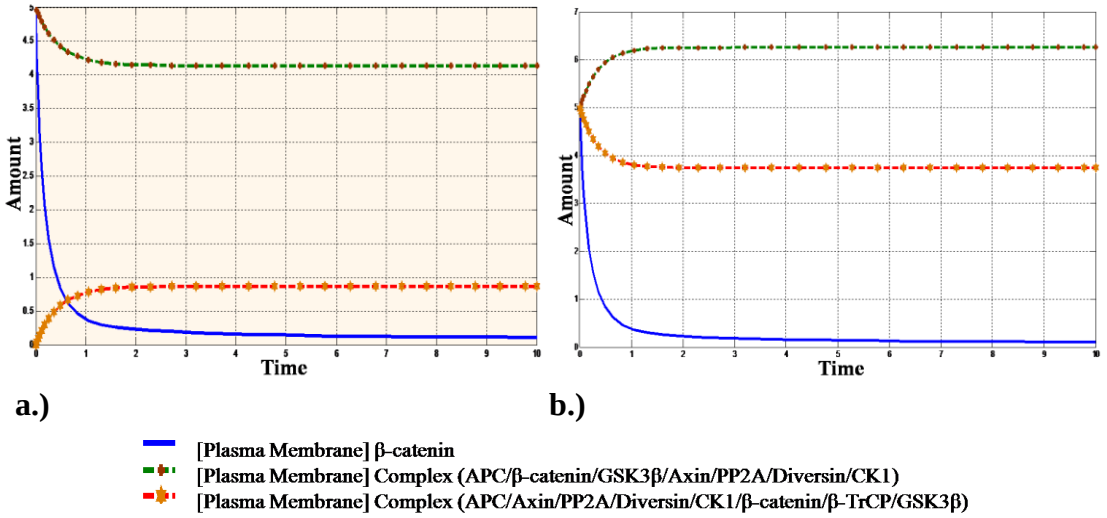


**Figure 5.** The dynamic behavior of components at varied timescale. (i.e. β-catenin (plasma membrane), Complex (APC, β-catenin, GSK3β, Axin, PP2A, Diversin, CK1 and Complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β); Concentration (Amount) in μM, Time in milliseconds.

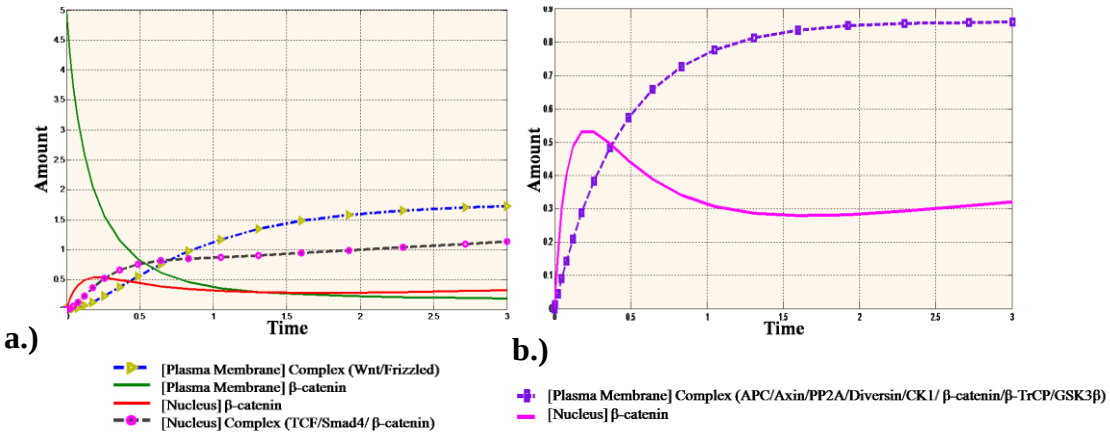


**Figure 6.** Dynamic behavior of components in different compartments. **6a.)** Behavior of the Complex (Wnt/Frizzled), β-catenin, and Complex (TCF, Smad4, β-catenin) and **6b.)** Behavior of the complex (i.e. β-catenin, and Complex (APC, Axin, PP2A, Diversin, CK 1, β-catenin, β-TrCP,

GSK3β)) in plasma membrane vs. the β-catenin in nucleus; Concentration (Amount) in μM, Time in milliseconds.

In the simulation (Figure 6a), we found that once the Wnt bind to the frizzled receptor the signaling initiates and the β-catenin level in the plasma membrane decreases and then rise up to certain extent (0.2 milliseconds) in the nucleus that leads to the transcription when binds to the TCF via tumor genes activation that leads a step towards the cancer progression. The next iteration was performed by considering the species (β-catenin, and Complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β)). Through the simulation, we found that the complex has the strong efficiency to suppress the level of the β-catenin in the nucleus and thus overcome the negative stimulation of the β-catenin (Figure 6b).

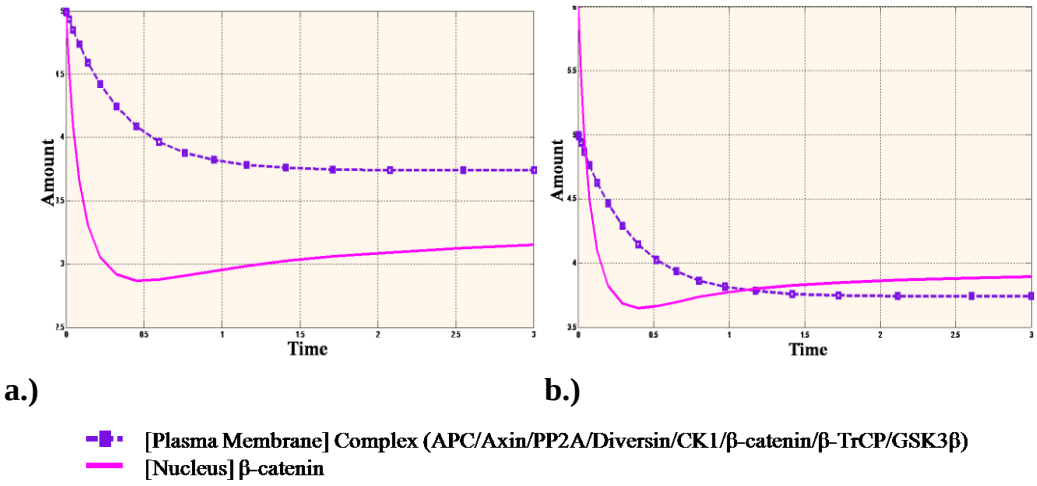


**Figure 7.** Dynamic behavior of components (i.e. β-catenin, and Complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β)); Concentration (Amount) in μM, Time in milliseconds.

The next iteration was performed by considering the species (β-catenin, and Complex (APC, Axin, PP2A, Diversin, CK1, β-catenin, β-TrCP, GSK3β)) but at varying concentrations. Two different simulations were run; Figure 7a is showing the overruling behavior of the complex over the β-catenin. In Figure 7b, when the concentration of the β-catenin is set high by one-unit, there is high possibility towards cancer progression. For the subgraph network detection, the network motifs were generated using statistical standards (with $Z > 2$; p-value $< 0.05$) through FANMOD tool [31]. From the significance profile, five subgraphs (3a, 4a, 5a, 6a, and 7a) were found to be overrepresented (Figure 8). The motif significance profiles illustrate that there is high difficulty

in recognizing the genes once the network start intensifying. Therefore, the overrepresented sub-graphs were annotated to determine the interacting gene partners. However there were multiple instances of the particular sub-graph type so we selected the ones with utmost frequency, this signifies the importance of these genes in the pathway (Figure 8). The overall five key genes were reflected i.e. Axin, APC, β-catenin, Lef1, and Myc with high statistical significance. Some additional significant components that could be important for in vitro and in vivo studies are MMP7, NLK and DVL1.

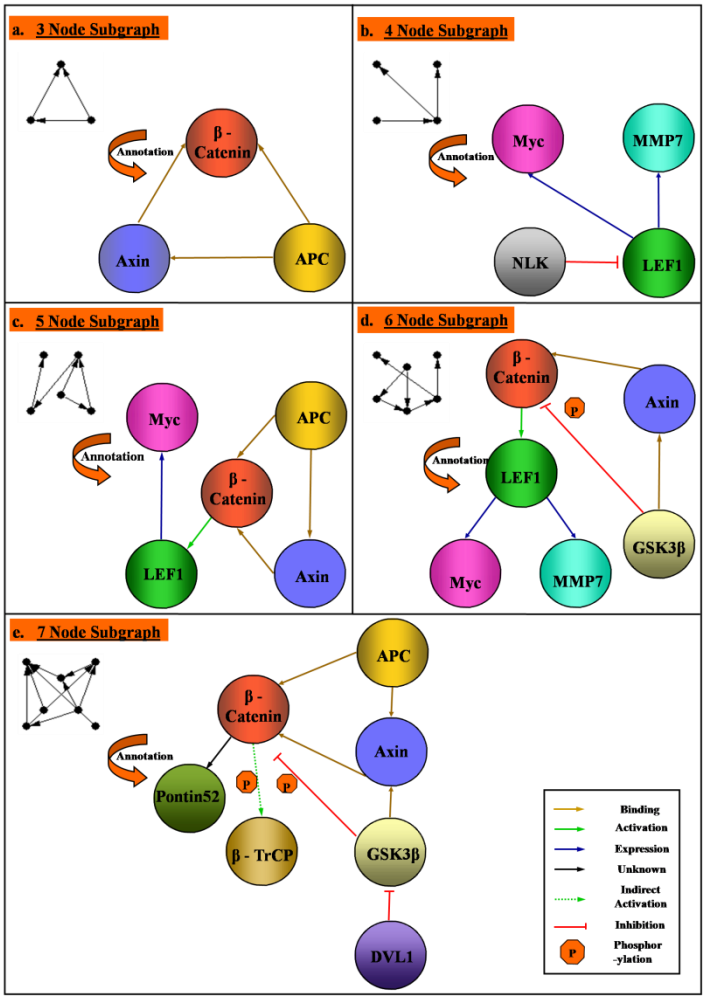


**Figure 8.** The sub-graph annotation of overrepresented motifs with their respective vital interactions

## CONCLUSION

The study illustrates the effect of the β-catenin and its regulation via different complexes that

include destruction complex that captures the β-catenin in the cytoplasm and prevent its transcription, and another one in the nucleus where it is bound to the transcription cell factor (TrCP) and made its progress towards CRC. The five key genes i.e. Axin, APC, β-catenin, Lef1, and Myc were detected to be putative regulatory elements of the pathway. Also, MMP7, NLK, and DVL1, are found to be the essential component of the study that can be possible putative elements for the disease progression study, thus should be considered for the experimental validation. We attempted to evaluate the current knowledge about the role of Wnt signaling pathway involved in colorectal carcinogenesis due to the target genes. Repeated patterns seems to play a significant role in the biological processes; through the subgraph study we came to know the interacting partners that shares very close neighborhood but their role in CRC is still unknown. If given attention these imprints can unravel the threads of the complex network and thus prove to be valuable insight for a lethal disease like CRC. Despite the advent strategies in drug development for Wnt pathway, the major hurdles in therapeutic intervention of the pathway still persist. With the aid of computerized modeling efforts, researchers can make sense of all the different bits and pieces of information that are accumulating at an ever-increasing pace. This will help the experimental biologists to understand the workings of living organisms with comprehensive models that enable them to connect the often disparate pieces of information.

## OBJECTIVE - III

### Study of Network and Structure Based Inference of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Impact on Colorectal Cancer

## METHODOLOGY

The network analysis followed the pathway signaling in terms of the quantitative simulations and determination of the network motifs that help to study the individual behavior of the pathway components to identify the putative markers. The structural variation study follows the effect of the damaging SNPs on the activity of the TGFβ1 protein and its progression towards CRC. The data collection of about 136 SNP's (missense, intronic variants etc.) for TGFβ1 gene were obtained through the dbSNP [27] and COSMIC databases [26]. Out of 136 SNP's final 37 were selected only with non-synonymous types. The overall methodology for performing the mentioned research is given in Figure 9.

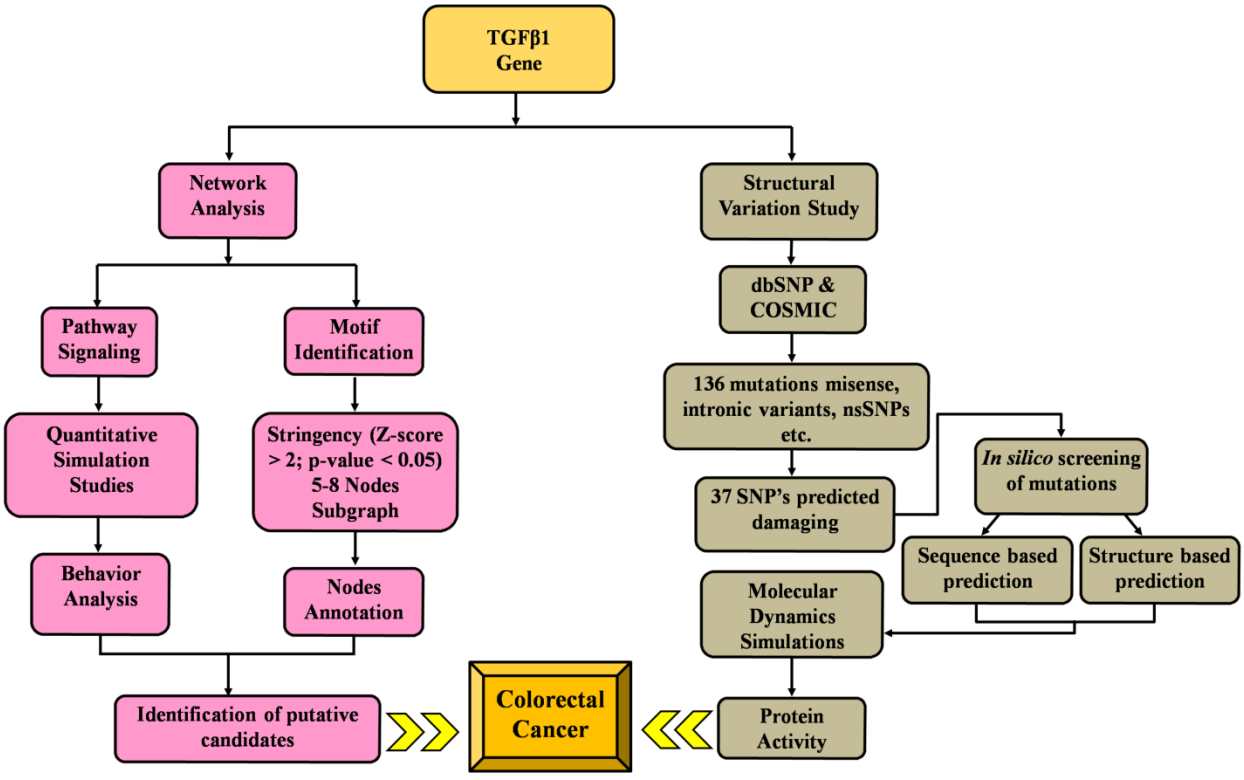


**Figure 9.** The flowchart of the procedure followed for analysis.

## RESULTS AND DISCUSSION

### Pathway Simulation

The TGFβ pathway simulations were performed through the Ode45 (Dormand-Prince) solver for all set of genes with varying parameters as illustrated in previous objective [29]. The parameters were selected based upon the standards given in the literature (Table 3) and absolute tolerance is set to 1.0E-6 with a relative tolerance of 0.0010. Depending on the function of the entities, separate simulations were run each time and at different time-period.

**Table 3.** The concentration values for the respective entities.

| Species | Type | Description | Initial Concentrations |
|---|---|---|---|
| S22 | COMPLEX | Total cellular homomeric S22 complexes | 0 |
| S24 | COMPLEX | Total cellular heteromeric S24 complexes | 0 |
| pS2tot | PHOSPHOPROTEIN | Total cellular pS2 | 0 |
| TGF | PROTEIN | TGFβ1 | 4 |
| R | PROTEIN | Nascent Receptors | 0.9 |
| S2C | PROTEIN | Cytoplasmic, unphosphorylated Smad2 | 1.2 |
| Rcom | PROTEIN | TGFβ bound receptors | 0.12 |
| pS2c | PHOSPHOPROTEIN | Total cytoplasmic pS2 | 0 |
| RcomS | PROTEIN | Mature, competent receptors | 0.05 |
| S2n | PROTEIN | Nuclear unphosphorylated Smad2 | 0.6 |
| S22n | COMPLEX | Nuclear homomeric S22 complexes | 0 |
| S4n | COMPLEX | Total nuclear Smad4 | 1 |
| S22c | COMPLEX | Cytoplasmic homomeric S22 complexes | 0 |
| pS2n | PHOSPHOPROTEIN | Total nuclear pS2 | 0 |
| pS2fn | PHOSPHOPROTEIN | Monomeric nuclear pS2 | 0 |
| S24n | COMPLEX | Nuclear heteromeric S24 complexes | 0 |
| S24c | COMPLEX | Cytoplasmic heteromeric S24 complexes | 0 |
| S4fc | PROTEIN | Monomeric cytoplamic Smad4 | 1 |
| S4c | PROTEIN | Total cytoplamic Smad4 | 1 |
| pS2fc | PHOSPHOPROTEIN | Monomeric cytoplasmic pS2 | 0 |
| S4fn | PROTEIN | Monomeric nuclear Smad4 | 1 |
| Rtot | GENE | Total receptors | 1 |
| RT | COMPLEX | Active receptors | 0 |
| RcomI | PROTEIN | - | 0.07 |
| Ract | PROTEIN | - | 0 |

Figure 10 shows the entities with significant behavior; the low concentration of the TGFβ1 while binding to the receptor (Figure 10a) activates the signaling nearby 4-units time-period at 0.03μM concentration. This states that even small amount of TGFβ1 protein is enough to initiate the pathway signaling. Once the receptor gets activated there is phosphorylation event of the Smad complex that leads to the complex stabilization (Figure 10b). Thus, inferring that the cytoplasm

signaling occurs at a fast pace. By changing the time from 50 μM to 5 μM, there is activation of the Smad complexes for first 5-units time period as the signal is consumed by the receptors (Figure 10b and 10c). This could be the crucial step, because if the complexes remain active for a long time then there is a chance that they will violate the gene regulation event and will overpower the process. When the concentration of the Smad4 is increased by 4-units the event is thought to be crucial as its overexpression has a tendency to perturb the cell cycle and initiate the disease progression (Figure 10d).

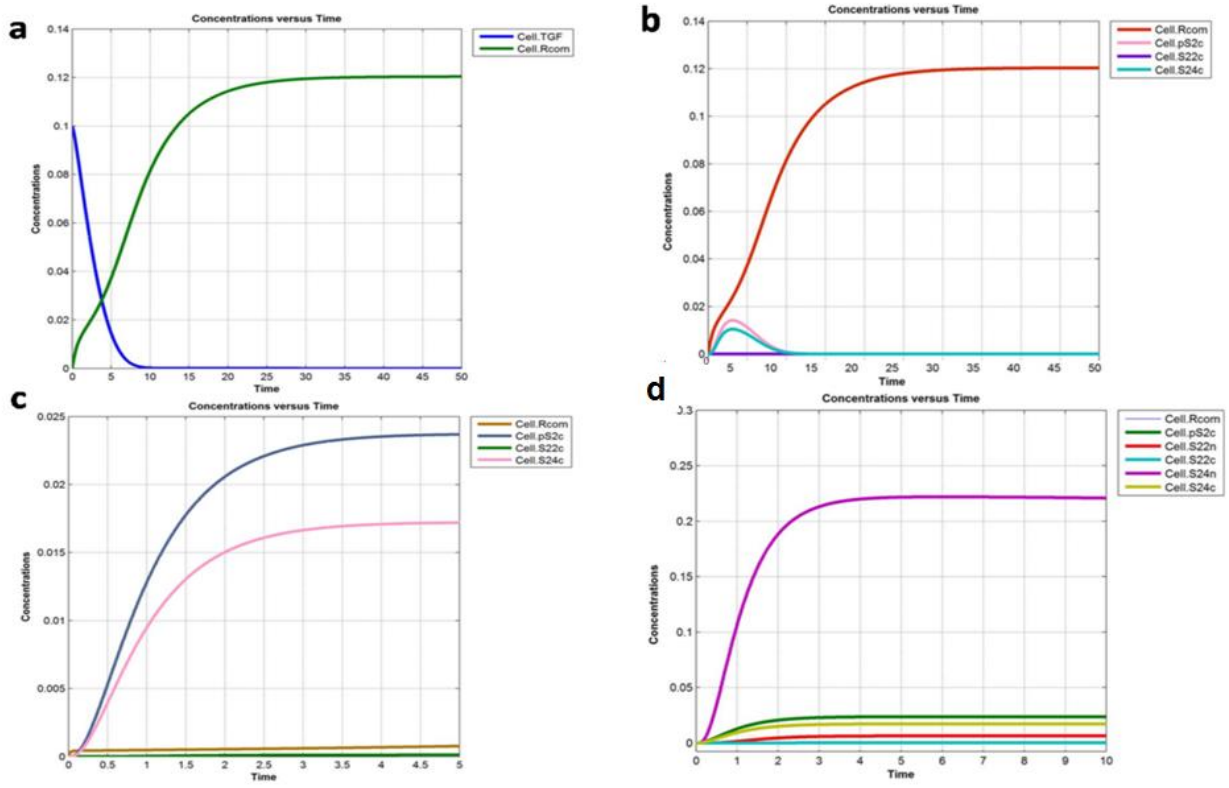


**Figure 10.** Quantitative simulations performed at different time-scale. a) Initiation of cell signaling via TGFβ1 b) Activation of the receptors and the Smad complexes c) Phosphorylation of the Smads d) Enhanced activity of the Smads; Concentration in μM, Time in milliseconds.

After, then regulatory patterns were determined in the TGFβ pathway to find the transcription factors involved. The retrieved sub-graphs were those containing 5-8 nodes and representative of them is given in Figure 10 a-d. In a 5-8 nodes sub-graphs, three variants of 3-node motifs and six variants of 4-node motifs were found to be overrepresented. These variants are TFs regulating the target gene, and the reprehensive SIMs (single input motifs). The disintegration of large sub-

graphs into small sub-graphs provided us several motifs matching with the motif-id's in the standard network motif dictionary provided by Alon et al [32]. The sub-graphs were annotated to determine the type of nodes that are significant to the pathway. The processes like binding, activation, expression, inhibition, and phosphorylation have been seen amongst the nodes of the sub-graphs. A large number of events have been captured through the sub-graphs, such as the role of Smad6 and INBHA in dual inhibition. The nodes like Smad4 and TNFα in rebuilding the network as evident from Figures 10 a-d. The TGFβR1, Smad1, Smad2, Smad4, Smad6 were found to be frequently occurring in almost all node of the sub-graphs, thus representing their key role in the pathway.

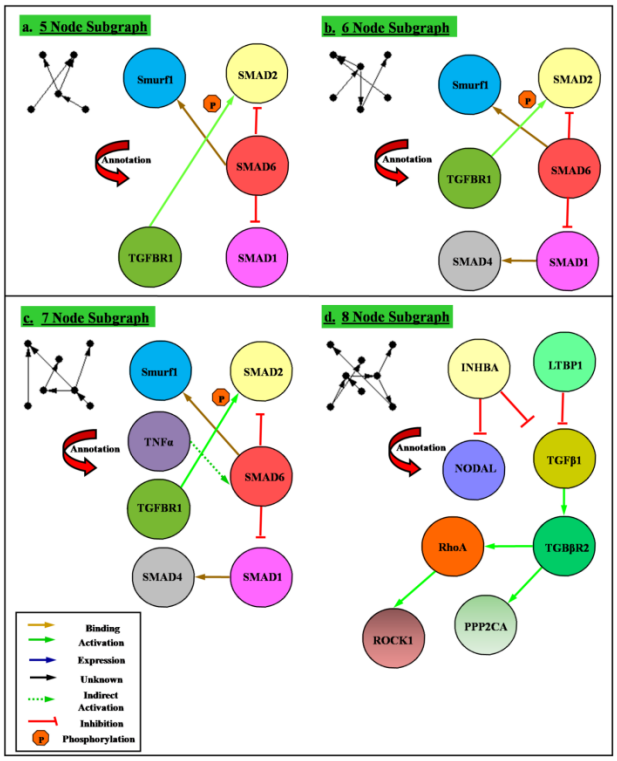


**Figure 11.** The over-represented sub-graphs annotated portray the crucial interactions.

Besides six key genes i.e. Smurf1, SMAD2, TGFβR1, SMAD1, SMAd6, and SMAd4 other new ones were determined such as TNFα, INHBA, LTBP1, TGFβ1, NODAL, PPP2CA, and ROCK1. The results postulate the computational prediction of the active sub-network that needs to be confirmed experimentally to determine their regulatory role in the disease conditions.

**Structural Analysis**

The structure analysis followed with the selection of the structure using Theseus-3D tool [33], that help to find the suitable structure from all PDB-hits (1KLA, 1KLC, 1KLD, 3KFD, and 4KV5) [20]. Using this method, 4KV5 was found out to be the median structure, this help to proceed with the further analysis. In first step, the active site of the protein was determined through CASTp [34]. The electrostatic potential elucidate the presence of mainly positive charges at the active site as blue color illustrates the positive charges, red color for negative charges and white for the neutral ones (Figure 12a). The results were confirmed through the CASTp as two pockets were selected (i.e. pocket 30 and pocket 29) based on the spatial position and relatively larger surface area. The pocket id 30 comprises of 13 residues from chain A (Tyr58, Leu62, Ser73, Ala75, Cys77, and Ser112) and chain B (Phe43, Cys44, Leu45, Cys77, Cys78, Val79, and Pro80) (Figure 12b shown as orange sticks). The residues that make up the pocket were principally all hydrophilic. In the case of the pocket id 29 the binding site comprised of 12 residues from chain A (Phe43, Cys44, Leu45, Cys78, Val79, and Pro80) and chain B (Tyr58, Leu62, Ser73, Ala75, Cys77, and Ser112) (Figure 12b shown as cyan sticks). We anticipate the binding of TGFβ1 to the receptor at the predicted cavities site, and we validated binding site activity by correlating it with the structure simulation studies.

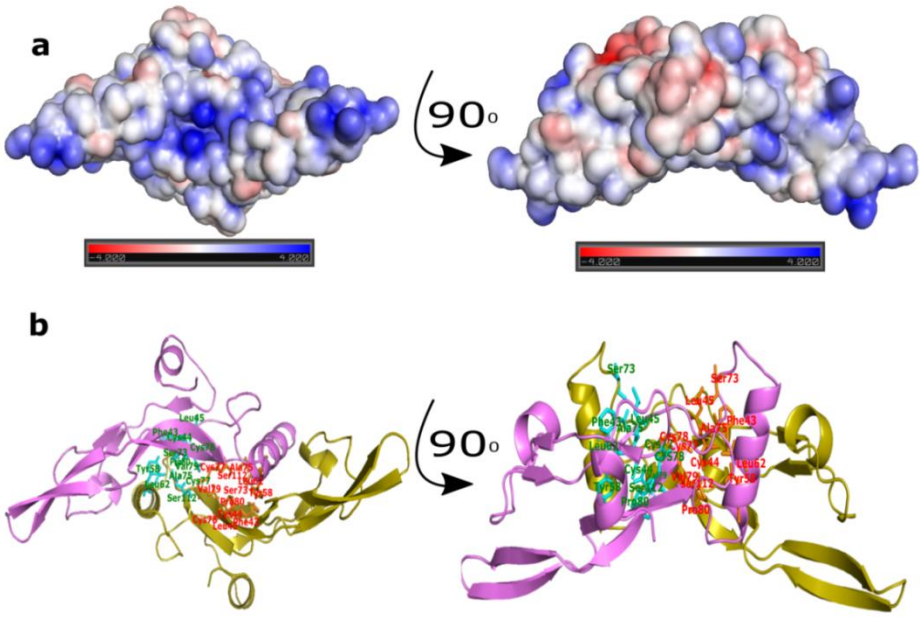


**Figure 12.** The binding pocket of protein 4kv5 obtained through CASTp results.

After, then molecular dynamics (MD) simulations were analyzed at global and local level. The significant changes were observed for the set of global parameters. Through RMSD analysis, the mutants L83R (2.4Å), N69Y (3.0Å), and L28F (3.1Å) were found to be similar to the native (2.9Å) trajectory up to 50ns, however after 50ns variations have been observed in comparison to the native trajectory, thus reflecting the structural shift. In comparison to the other mutants L28F and G46R, have shown major variations. Through RMSF significant changes were found in comparison to the native one. The mutants L28F, N69Y, and L83R were shown to be highly fluctuating for the residues Arg94, and Tyr50. The Rg measure the stability index of the protein structure that undergoes mutation or denaturation. It was measured to be 2.12nm and 2.14nm respectively for the mutants. The SASA is one of the major contributing factors for protein stability as it predicts the conformational changes to the proteins upon binding to the substrate. Also, the superposition of frames, i.e. the first timeframe (0ns) to the last timeframe (100ns) reflected the conformational changes in terms of secondary structure specifically at the loop region, helix region and also at the beta-sheet region (Figure 13a). Again mapping was performed by considering the residues around the 5Å regions of the residues Phe28, Arg46, Tyr69, and Arg83. The structural mapping states the significant structural differences in the residues within a 5Å radius of mutation for the first and last time frames (Figure 13 b-e).

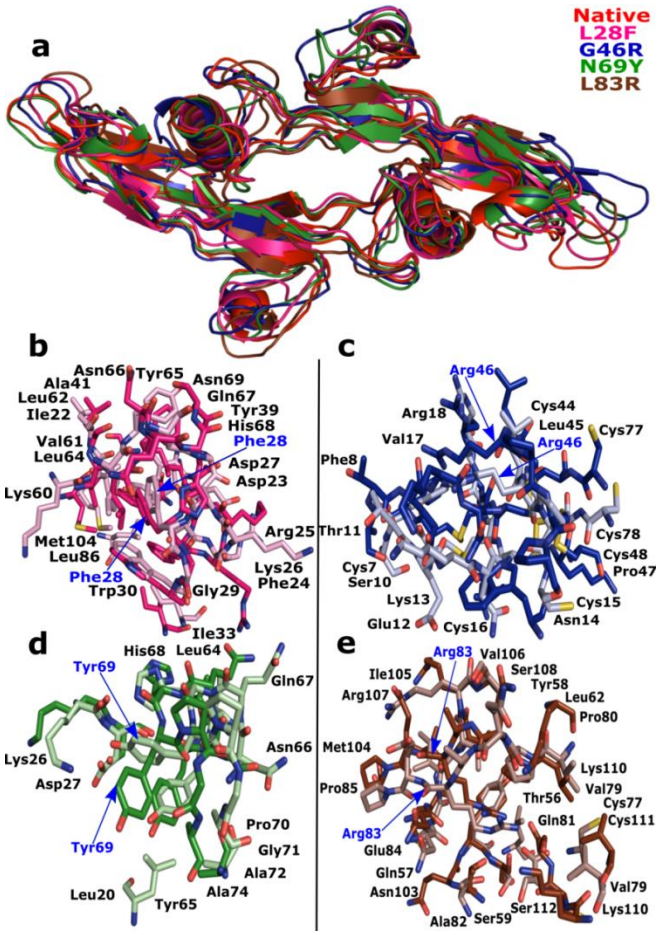


**Figure 13.** The comparisons of the first and last frame of the native and mutant structures obtained through molecular dynamics. Mutant residues are being indicated through blue arrows (b-e).

The polar interactions determined by Ligplot were cross-validated using the H-bond command of the GROMACS. Simultaneously we looked at some hydrogen bonds throughout the trajectory. After the cross-check, it has been found that certain bonds were formed at the beginning of the simulation and maintained throughout the 100 ns trajectory, while other were formed at the end of the simulation only. This concludes that the former bonds were formed via structural adjustments at internal level thus compensating the entropic cost of a mutating residue by stabilizing the dimer structure. The bonds found only in the last timeframe of the trajectory (at 100ns) are transient bonds that are formed due to the minor fluctuations and not contributing to the mutant structure stability (Figure 14).

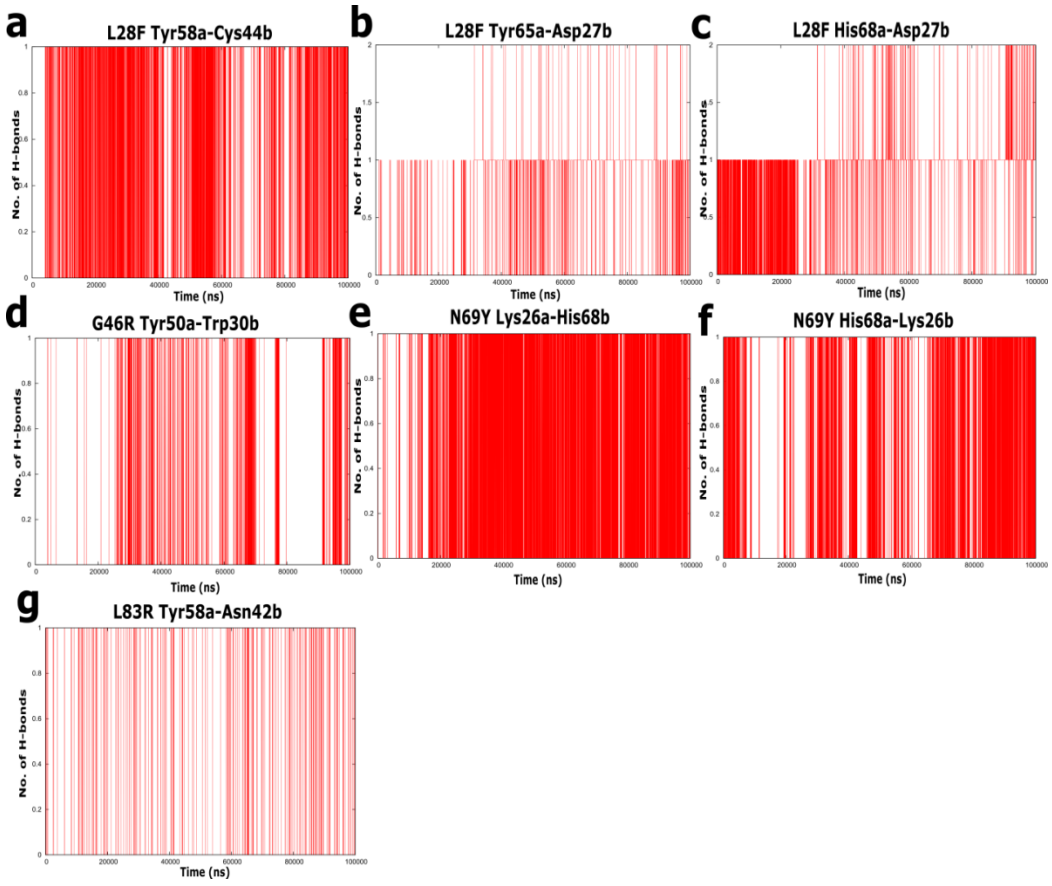


**Figure 14.** Polar-interactions in respective time-frame.

Next, the VMD tool was used to visualize the trajectories for all 10000 frames at 100ns for the native and mutants [35]. It illustrated variations of the structure at different time frames through localized RMSD differences, as the red region was displaying the regions with high deviation, green regions with intermediate deviations, and blue ones with slightest deviations (Figure 15). This shows that although no major changes are happening to the overall protein structure, the deviation has been found at the secondary structure regions (i.e. loop region, helix region) that have chances of structure conformational change thus altering functionality.

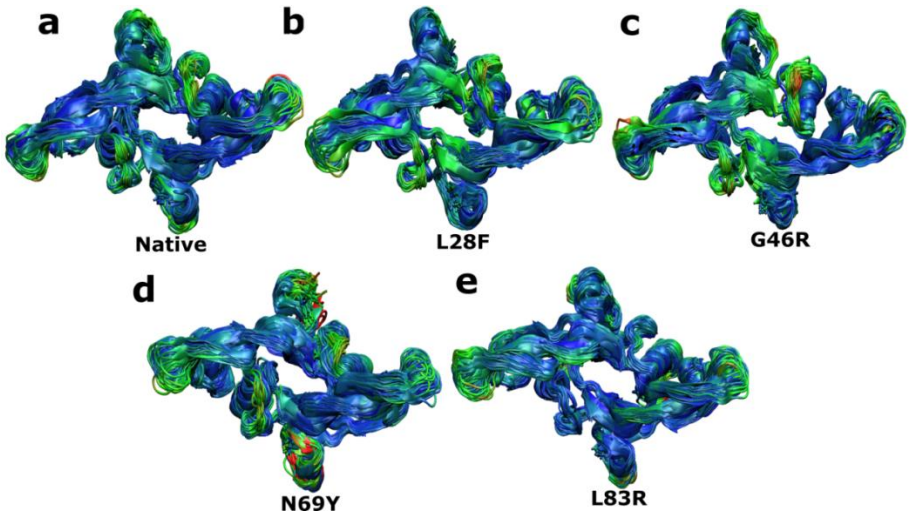


**Figure 15.** The superposed frames snapshots of 100ns simulation of TGFβ1 dimer site mutants.

## CONCLUSION

It is anticipated that this pathway and structure level study will give a new insight into the computational and experimental biologists for designing therapeutics for the CRC. The network analysis identified six key genes i.e. Smurf1, SMAD2, TGFβR1, SMAD1, SMAD6, and SMAD4 regulating the TGFβ pathway along with the new regulatory elements such as TNFα, INHBA, LTBP1, TGFβ1, NODAL, PPP2CA, and ROCK1 that were found to involved in the major processes in a pathway and thus should be considered experimentally to validate their role in CRC. Thus computer simulations can quickly investigate different experimental conditions for the biological system of interest, and through the analysis, only the most relevant cases can be assessed easily thus saving time and money. From the structural study, the SNP's rs199946261, rs768250306, rs763943753, and rs541829714 were found to be the damaging and have a potential to hinder the functionality of the protein. We hope that this computational analysis will provide a broad view of the disease to the biological researchers so that effective methods can be developed to eradicate this disease. The results presented here will be helpful for an experimental biologist to test the effect of SNPs and the subsequent phenotypical response of the mutations.

The study has implicated the critical role of DNA repair mechanisms in Lynch Syndrome associated malignancies that is performed through computational means. Both top-down and bottom-up approaches have been utilized for deciphering candidate markers for CRC and Lynch syndrome allied cancer forms. Various biomolecules such as genes, TFs, proteins, vital

regulatory elements and interactions among them have been elucidated at systems level for better understanding of the bioprocesses concerned with DNA repair. The outcome of the research would certainly assist to further carry out the research with additional known factors and the regulatory mechanisms. Eventually, the exhaustive mechanistic perspective of the DNA repair mechanisms and its impact on these cancer types will help in finding the advanced knowledge of the cancers and thus to apply the correct strategies to overcome them. Therefore, the work performed will provide innovative paradigms for genetic susceptibility, prevention, diagnosis at an earliest condition.

## FUTURE PROSPECTS

- ✓ It is anticipated that this web based comprehensive resource and would serve as a valuable accompaniment for retrieving crucial information for the CRC, endometrial cancer, and Lynch syndrome. The database will be updated on a regular basis to keep updated information to the academicians and researchers that help to perform the research in right direction.
- ✓ The proposed five key genes (Axin, APC, β-catenin, Lef1, and Myc) in the Wnt signaling pathway can be experimentally verified and the consequences of their impact on CRC could be observed. Identification of these key genes and the manner they behave during cancer progression can make them plausible biomarkers for the disease condition.
- ✓ The projected key genes of the TGFβ pathway and the four potential damaging SNPs candidates (rs199946261, rs768250306, rs763943753, and rs541829714) may be authenticated experimentally thus saving time and resources; contributing to timely systematic understanding of CRC. This will aid in fast drug discovery process and help biologist to take necessary actions to target the appropriate region of a protein.

## References


[1] A. Balestrazzi, V. M. Achary, A. Macovei, K. O. Yoshiyama, and A. N. Sakamoto, "Editorial: Maintenance of Genome Integrity: DNA Damage Sensing, Signaling, Repair, and Replication in Plants," *Front Plant Sci*, vol. 7, p. 64, 2016.

[2] E. C. Friedberg, G. C. Walker, W. Siede, and R. D. Wood, *DNA repair and mutagenesis*: American Society for Microbiology Press, 2005.

[3] M. Papamichos-Chronakis and C. L. Peterson, "Chromatin and the genome integrity network," *Nat Rev Genet*, vol. 14, pp. 62-75, Jan 2013.

[4] A. Livnat, "Interaction-based evolution: how natural selection and nonrandom mutation work together," *Biology direct,* vol. 8, p. 24, 2013.
[5] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "DNA damage and repair and their role in carcinogenesis," 2000.
[6] G. M. Cooper and R. E. Hausman, *The cell: Molecular approach*: Medicinska naklada, 2004.
[7] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, "Making sense of cancer genomic data," *Genes Dev,* vol. 25, pp. 534-55, Mar 15 2011.
[8] M. C. Poirier, "Chemical-induced DNA damage and human cancer risk," *Discov Med,* vol. 14, pp. 283-8, Oct 2012.
[9] P. S. Patil, A. Saklani, P. Gambhire, S. Mehta, R. Engineer, A. De'Souza, *et al.*, "Colorectal Cancer in India: An Audit from a Tertiary Center in a Low Prevalence Area," *Indian J Surg Oncol,* vol. 8, pp. 484-490, Dec 2017.
[10] G. Poulogiannis, I. M. Frayling, and M. J. Arends, "DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome," *Histopathology,* vol. 56, pp. 167-79, Jan 2010.
[11] R. M. Byrne and V. L. Tsikitis, "Colorectal polyposis and inherited colorectal cancer syndromes," *Ann Gastroenterol,* vol. 31, pp. 24-34, Jan-Feb 2018.
[12] R. Gryfe, "Inherited colorectal cancer syndromes," *Clin Colon Rectal Surg,* vol. 22, pp. 198-208, Nov 2009.
[13] Q. Wang, "Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes," *Acta Pharmacol Sin,* vol. 37, pp. 143-9, Feb 2016.
[14] R. L. Siegel, K. D. Miller, S. A. Fedewa, D. J. Ahnen, R. G. S. Meester, A. Barzi, *et al.*, "Colorectal cancer statistics, 2017," *CA Cancer J Clin,* vol. 67, pp. 177-193, May 6 2017.
[15] T. Patra, S. Mandal, N. Alam, and N. Murmu, "Clinicopathological trends of colorectal carcinoma patients in a tertiary cancer centre in Eastern India," *Clinical Epidemiology and Global Health,* 2017.
[16] K. Tariq and K. Ghias, "Colorectal cancer carcinogenesis: a review of mechanisms," *Cancer Biol Med,* vol. 13, pp. 120-35, Mar 2016.
[17] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research,* vol. 32, pp. 1792-1797, 2004.
[18] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology,* vol. 215, pp. 403-410, 1990.
[19] M. Magrane and C. UniProt, "UniProt Knowledgebase: a hub of integrated protein data," *Database (Oxford),* vol. 2011, p. bar009, 2011.
[20] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, *et al.*, "Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules," *Acta Crystallographica Section D: Biological Crystallography,* vol. 54, pp. 1078-1084, 1998.
[21] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright, "miRBase: microRNA sequences, targets and gene nomenclature," *Nucleic Acids Res,* vol. 34, pp. D140-4, Jan 1 2006.
[22] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research,* vol. 28, pp. 27-30, 2000.

[23] C. v. Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, "STRING: a database of predicted functional associations between proteins," *Nucleic acids research*, vol. 31, pp. 258-261, 2003.
[24] M. Ongenaert, L. Van Neste, T. De Meyer, G. Menschaert, S. Bekaert, and W. Van Criekinge, "PubMeth: a cancer methylation database combining text-mining and expert annotation," *Nucleic acids research*, vol. 36, pp. D842-D846, 2007.
[25] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic acids research*, vol. 37, pp. W623-W633, 2009.
[26] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, *et al.*, "COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer," *Nucleic acids research*, vol. 39, pp. D945-D950, 2011.
[27] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, *et al.*, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, pp. 308-311, 2001.
[28] T. MathWorks, "MATLAB and Statistics Toolbox Release 2012a,", *The MathWorks, Inc., Natick, Massachusetts, United States.*, 2012.
[29] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of computational and applied mathematics*, vol. 6, pp. 19-26, 1980.
[30] A. Drager, N. Hassis, J. Supper, A. Schroder, and A. Zell, "SBMLsqueezer: a CellDesigner plug-in to generate kinetic rate equations for biochemical networks," *BMC Syst Biol*, vol. 2, p. 39, Apr 30 2008.
[31] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152-1153, 2006.
[32] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824-827, 2002.
[33] D. L. Theobald and P. A. Steindel, "Optimal simultaneous superpositioning of multiple structures with missing data," *Bioinformatics*, vol. 28, pp. 1972-9, Aug 01 2012.
[34] J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang, "CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues," *Nucleic acids research*, vol. 34, pp. W116-W118, 2006.
[35] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *J Mol Graph*, vol. 14, pp. 33-8, 27-8, Feb 1996.

# PUBLICATIONS

**Research Publications:**

- **Ankita Shukla**, Manika Sehgal, Tiratha Raj Singh (2015). Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives. ***Gene. 2015 Jun 15; 564(2):109-118. Scopus, IF: 2.3***
- **Ankita Shukla**, Ahmed Moussa, and Tiratha Raj Singh. DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies. ***PLoS One. 2016 Jun 8; 11(6):e0157031. Scopus, IF: 3.3***
- **Ankita Shukla**, Ragothaman Yennamali, and Tiratha Raj Singh. Study of Network and Structure Based Inference of Functional Single Nucleotide Polymorphisms of TGFβ1 Gene and its Impact on Colorectal Cancer (CRC). **GeneReports.*, 2018 Jun ,11;131-142. Scopus.***
- **Ankita Shukla**, and Tiratha Raj Singh. Network-Based Approach to Study Dynamics of Wnt Pathway Regulatory Elements in Colorectal Cancer. **(Under-review).**

**Conference Publications:**

- **Ankita Shukla**, Manika Sehgal, Tiratha Raj Singh (2014). Hydroxymethylation and its Role in DNA Repair Associated Disease. ***Proceedings of the International Conference on Life Sciences, Informatics, Food and Environment [Noida, India: 29-30, August 2014].***
- **Ankita Shukla**, Ashwani Kumar, Manika Sehgal, Tiratha Raj Singh (2015). Modeling and Simulation of Colorectal Cancer Pathways: An Insight into Regulatory Elements and Biological Processes. ***Proceedings of the Annual International Conference on Advances in Biotechnology [5th: Kanpur, India: 13-15 March 2015], pp.85-89.***
- **Ankita Shukla**, Ragothaman Yennamalli, Tiratha Raj Singh (2017). Structure Based Inference of Functional Single Nucleotide Polymorphism "L28F" and to Determine its Role in TGFβ1 Allied Colorectal Cancer (CRC). ***Proceedings of the INBIX-2017, in International Journal of Bioinformatics Research and Applications*.**

**Book Chapters**

- **Ankita Shukla** and Tiratha Raj Singh. "Computational Network Approaches and Their Applications for Complex Diseases." Translational Bioinformatics and Its Application. **Springer Netherlands, 2017. 337-352. *SCOPUS Indexed*.**
- Tiratha Raj Singh, **Ankita Shukla**, Bensellak Taoufik Ahmed Moussa and Brigitte Vannier. "Metabolome Analysis: A Disease Biomarker Discovery". **(Under Revision).**

**General Publications**

- Tiratha Raj Singh, **Ankita Shukla** (2017). Bioinformatics to Systems Biology: A Journey of Knowledge Discovery. CSI Communications, 41 (4), 10-13.